G10L 3/02

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

Recherche

(11) EP 0 865 026 A2

(12) EUROPÄISCHE PATENTANMELDUNG

(43) Veröffentlichungstag:
16.09.1998 Patentblatt 1998/38

(51) Int. Cl.[6]: **G10L 3/02**

(21) Anmeldenummer: **98104455.5**

(22) Anmeldetag: **12.03.1998**

(84) Benannte Vertragsstaaten:
AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Benannte Erstreckungsstaaten:
AL LT LV MK RO SI

(30) Priorität: **14.03.1997 DE 19710545**

(71) Anmelder:
GRUNDIG Aktiengesellschaft
90762 Fürth (DE)

(72) Erfinder: **Carl, Holger, Dr.**
90762 Fürth (DE)

(54) Effizientes Verfahren zur Geschwindigkeitsmodifikation von Sprachsignalen

(57) Die Erfindung betrifft ein Verfahren zur Geschwindigkeitsmodifikation von Sprachsignalen, insbesondere digitalisierten Sprachsignalen. Bei diesem Verfahren wird ein analoges Sprachsignal digitalisiert und in einem Speicher gespeichert. Außerdem wird ein Faktor α definiert, um den das Sprachsignal verlängert oder verkürzt wird. Über das Sprachsignal wird eine Fensterfunktion mit einem ersten steigenden Abschnitt, einem zweiten, sich direkt an den ersten Abschnitt anschließenden, konstanten Abschnitt und einem dritten, sich direkt an den zweiten Abschnitt anschließenden, fallenden Abschnitt, gelegt.

EP 0 865 026 A2

Printed by Xerox (UK) Business Services
2.16.3/3.4

Beschreibung

Gegenstand der Erfindung ist ein Verfahren zur Geschwindigkeitsmodifikation von Sprachsignalen im Zeitbereich, insbesondere eine effiziente Overlap-Add-Methode.

In verschiedenen Bereichen der Verarbeitung von Sprach- und Audiosignalen ist eine Veränderung der Wiedergabegeschwindigkeit dieser Signale erwünscht, möglichst ohne daß damit eine Beeinträchtigung ihrer Natürlichkeit und - im Fall von Sprache - ihrer Verständlichkeit verbunden wäre. Dieses Ziel, den Klangcharakter zu erhalten, kann man aus technischer Sicht folgendermaßen formulieren: Trotz einer Modifikation der Zeitskala dieser Signale sollen ihre Kurzzeitspektraleigenschaften unverändert bleiben. Insbesondere bedeutet das für Sprachsignale, daß Grundfrequenz und Formanten bei der Geschwindigkeitsmodifikation erhalten bleiben müssen.

Die Zeitstauchung oder Zeitdehnung von Audiosignalen wird in Studios eingesetzt, zum Beispiel mit dem Ziel, Werbesendungen auf die vorgesehene Länge zu trimmen. Auch in der Diktiertechnik ist die Anpassung der Wiedergabegeschwindigkeit an die Bedürfnisse bzw. Fähigkeiten der Schreibkraft von Bedeutung. Eine weitere Anwendung besteht bei der Echtzeitübertragung von Sprachsignalen, bei der Datenpakete mit variabler Verzögerung beim Empfänger eintreffen. Durch Anwendung der Geschwindigkeitsmodifikation kann man hier die Über-Alles-Verzögerung im Mittel geringer halten als das Worst-Case Delay der Übertragungsstrecke, ohne daß ein zu spät eintreffendes Datenpaket zu Aussetzern oder anderen, ähnlich störenden Effekten führen würde.

Für viele Anwendungen ergeben sich neben dem Wunsch nach möglichst hoher Klangqualität die folgenden zusätzlichen Anforderungen an das Verfahren:

Eine kostengünstige Echtzeitrealisierung muß erzielbar sein, und es muß zur Laufzeit eine nach Möglichkeit stufenlose Änderung des Geschwindigkeitsmodifikationsfaktors möglich sein. Von Vorteil ist ohne Zweifel auch, wenn der Algorithmus ohne eine stets fehlerbehaftete Pitch-Schätzung auskommt.

Aus "Method for Time or Frequency Compression-Expansion of Speed", von G. Fairbaks und R. P. Jaeger, Inst. of Radio Engineers Trans. on Audio, Vol. AU-2, No. 1 pp. 7-12, Jan. 1954, sind erste Untersuchungen zur Sprachsignalstauchung bzw. Sprachsignaldehnung bekannt. Häufig wurden seitdem Frequenzbereichsverfahren eingesetzt - naheliegend, da, wie eingangs erwähnt, die Kurzzeitspektraleigenschaften des Sprachsignals erhalten bleiben sollen. Seit Mitte der achtziger Jahre sind vergleichsweise einfache im Zeitbereich arbeitende Overlap-Add-Verfahren bekannt, mit denen sehr gut klingende zeitskalierte Sprachsignale erzeugt werden können.

In "Signal Estimation from Modified Short-Time Fourier Transform", von D. W. Griffin, in IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-32, No. 2, pp. 236-242, Apr. 1984, berichten Griffin und Lim von Experimenten mit einer sehr aufwendigen iterativ arbeitenden Phasenbestimmung. Auf diesen Ansatz nimmt wiederum die Veröffentlichung von S. Roucos und A. M. Wilgus "High Quality Time-Scale Modification for Speech", IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 493-496, 1985, Bezug, die eine Zeitbereichsmethode vorschlagen, die mittels eines Overlap-Add-Ansatzes zeitskalierte Sprachsignale erzeugt. Bei diesem sogenannten SOLA-Verfahren (SOLA = Synchronized OverLap-Add) erfolgt eine Synchronisation der in regelmäßigen Abständen dem Originalsignal entnommenen Abschnitte durch Verschiebung vor der jeweils entsprechenden Fensterung und Addition im Zielsignal. Dies entspricht im weiteren Sinne der Phasenoptimierung, wie sie in den Frequenzbereichsverfahren durchgeführt wird. Eng mit dem SOLA-Algorithmus verwandt ist das sogenannte WSOLA-Verfahren (WSOLA = Waveform Similarity OverLap-Add), das W. Verhelst und M. Roelands in "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech", IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 554-557, 1993, und "Waveform Similarity Based Overlap-Add (WSOLA) for Time-Scale Modification of Speech: Structures and Evaluation", Int. Conf. on Speech Communication and Technology, pp. 337-340, 1993, vorstellen. Der Hauptunterschied zwischen diesen beiden Ansätzen besteht in der Synchronisation, die im WSOLA-Verfahren durch versetztes Entnehmen von Segmenten aus dem Originalsignal durchgeführt wird, was sich gegenüber dem SOLA-Prinzip vor allem aufwandsmindernd auswirkt.

7

Aufgabe der Erfindung ist es, ein Verfahren zur Geschwindigkeitsmodifikation von Sprachsignalen im Zeitbereich anzugeben, das besonders effizient arbeitet und gegenüber dem Stand der Technik weniger Aufwand erfordert.

Diese Aufgabe wird durch die Merkmale der Ansprüche 1 und 2 gelöst. Vorteilhafte Ausgestaltungen der Erfindung sind in der nachfolgenden Beschreibung angegeben.

Die Erzeugung der mit dem Faktor α zeitskalierten Version y(k) eines Sprachsignals x(k) erfolgt gemäß der Synthese

$$y(k)=\sum_{\lambda=-\infty}^{\infty}(k+\lambda(\alpha\text{-}1)L+\Delta_\lambda)\,w(k-\lambda L)$$

mit einer Fensterfunktion

$$w(k)=\begin{cases} v(k) & \text{für} \quad 0<k<N \\ 1 & \text{für} \quad N\le k<L \\ 1-v(k-L) & \text{für} \quad L\le k<L+N \\ 0 & \text{sonst} \end{cases}$$

Die hierin vorkommende für k=0, ... , N-1 definierte Funktion v(k) ist dabei sinnvollerweise zwischen ihren Extrema $v(0)=\varepsilon_0$ mit $0<\varepsilon_0<<1$ und $v(N-1)=1-\varepsilon_1$ mit $0<\varepsilon_1<<1$ monotonwachsend.

Die angegebene w(k)-Definition stellt sicher, daß die für sinnvolles Overlap-Add notwendige Bedingung

$$\sum_{\lambda=-\infty}^{\infty} w(k-\lambda L) = 1 \;\forall k \in \{-\infty,\ldots,\infty\}$$

erfüllt ist.

Die in obiger Synthesegleichung enthaltene Verschiebevariable Δ_λ ist zwecks der erwähnten Synchronisation aus einem "Toleranzbereich" $-\Delta_{max}, \ldots, \Delta_{max}$ zu bestimmen.

Die prinzipielle Vorgehensweise ist wie folgt:

Aus dem Originalsignal x(k) werden in - abgesehen von einem synchronisationsbedingten "Jitter" - regelmäßigen αL Werte betragenden Abständen Segmente der Länge L+N entnommen und nach Gewichtung mit w(k) jeweils um L Abtastwerte versetzt aufaddiert. Das auf diese Weise erhaltene Signal y(k) ist gegenüber x(k) um den Faktor α beschleunigt, das heißt, daß eine im Originalsignal x(k) enthaltene Äußerung von K Abtastwerten Länge durch dieses Vorgehen auf einen y(k) -Abschnitt der Länge K/α abgebildet, also verkürzt und damit in der Wiedergabe beschleunigt für $\alpha > 1$, bzw. verlängert, das heißt verlangsamt, wird, wenn $\alpha < 1$ ist.

Die Synchronisation der zu überlappenden Abschnitte ist für die resultierende Klangqualität von großer Bedeutung. Hierzu wird der folgende Ansatz verwendet: Während der Abarbeitung des Verfahrens kann zu jedem dem Signal x(k) entnommenen Segment für den nächsten Schritt als "Idealsegment" der um L Abtastwerte versetzte Abschnitt von x(k) angesehen werden, da durch diese Wahl die Overlap-Add-Operation wieder das Originalsignal x(k) reproduzieren würde. Die erwünschte Zeitskalierung erfordert nun aber, daß für die Overlap-Add-Synthese i. a. ein anderer, gegenüber dem "Idealsegment" versetzter Abschnitt von x(k) ausgewählt wird. Die bestmögliche Synchronisation ist gegeben, wenn der für die Overlap-Add-Operation benutzte Abschnitt größtmögliche Ähnlichkeit ("Waveform Similarity") mit dem "Idealsegment" aufweist.

Als Kriterium für die Ähnlichkeit der genannten Segmente bieten sich verschiedene Maße an. Naheliegend ist beispielsweise die Benutzung des Korrelationskoeffizienten. Während W. Verhelst und M. Roelands in "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speed", in IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 554-557, 1993, und "Waveform Similarity Based Overlap-Add (WSOLA) for Time-Scale Modification of Speech: Structures and Evaluation" in Int. Conf. on Speech Communication and Technology, pp. 337-340, 1993, für die Auswertung des Ähnlichkeitsmaßes das komplette Segment der Länge L+N herangezogen haben, erscheint es als vollkommen ausreichend, die Berechnung auf den Bereich der N Abtastwerte zu beschränken, in dem die Segmente tatsächlich überlappen.

Für die weiteren Darstellungen ist es hilfreich, die folgende Vektornotation einzuführen:

Der N Werte lange Abschnitt des "Idealsegments", in dem die Überlappung mit dem neu zu bestimmenden Segment stattfinden wird, sei mit x bezeichnet, die ersten N Werte des verschobenen Segments mit x_q. Die Gewichtung dieses Abschnitts mit der steigenden Flanke des Fensters wird durch Multiplikation dieses Vektors mit einer Diagonalmatrix V repräsentiert, die mit den Werten v(0), ... , v(N-1) besetzt ist. Entsprechend wird die Gewichtung des Idealsegmentabschnitts x mit der fallenden Flanke des Fensters durch Multiplikation mit **1** - V dargestellt, wobei **1** die N×N-Einheitsmatrix bezeichnet. Der im kritischen Überlappungsbereich aus der Overlap-Add-Synthese resultierende y(k)-Abschnitt lautet damit

$$y=(1-V)x + \bar{V}x_q$$

Beispielsweise läßt sich nun als Maß für die Ähnlichkeit der hierbei beteiligten Komponenten eine Kreuzkorreliertenberechnung gemäß

$$C_\delta = x^T(1-V)^T V x_q$$

angeben. Die Maximierung dieses Ausdrucks bezüglich der sich in x_q wiederfindenden Verschiebung $\delta \in \{-\Delta_{max}, \dots, \Delta_{max}\}$ liefert die für das betrachtete Segment im Sinne des angesetzten Ähnlichkeitsmaßes optimale Verschiebung Δ_λ.

Die Berechnung der C_δ erfordert alle L Abtastwerte 2N Multiplikationen für die Vorabberechnung des Ausdrucks $x^T(1-V)^TV$ sowie anschließend $(2\Delta_{max}+1)N$ Multiplikationen und Additionen.

Dies stellt gegenüber W. Verhelst und M. Roelands in "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speed", in IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 554-557, 1993, und "Waveform Similarity Based Overlap-Add (WSOLA) for Time-Scale Modification of Speech: Structures and Evaluation" in Int. Conf. on Speech Communication and Technology, pp. 337-340, 1993, eine Aufwandsreduktion um den Faktor zwei dar, der sich für L>N sogar noch erhöht. Die Beschränkung der Ähnlichkeitsberechnung auf den Bereich der Überlappung hat keinerlei negative Auswirkungen auf die Qualität der zeitskalierten Sprachproben.

Ein anderer Ansatz für die Synchronisation ist, anstelle der Maximierung der "Waveform Similarity" den Fehler zwischen dem synthetisierten Signal y und dem Originalsignal x zu minimieren. Eine einfache willkürliche Wahl ist, für diesen Fehler den quadratischen Ausdruck

$$E_\delta = \|x-y\|^2$$

anzusetzen.

Bei Vernachlässigung der Vorabberechnungen beläuft sich der für die Auswertung von E_δ anfallende Aufwand auf $(2\Delta_{max}+1)4N$ DSP-Operationen alle L Abtastwerte. Hierunter werden solche Operationen verstanden, die ein Signalprozessor mit gängiger Architektur in einem Schritt abarbeiten kann.

Ein weiterer Ansatz besteht darin, anstelle des absoluten Fehlers den relativen Fehler

$$R\partial = \frac{\|x - y\|^2}{\|y\|^2}$$

zu minimieren, was als SNR-Maximierung interpretiert werden kann. $(2\Delta_{max}+1)5N$ Operationen sind hier vor jeder Overlap-Add-Operation erforderlich.

Patentansprüche

1. Verfahren zur Geschwindigkeitsmodifikation von Sprachsignalen, insbesondere digitalisierten Sprachsignalen, bei dem

   - ein analoges Sprachsignal digitalisiert wird, wodurch ein digitalisiertes Sprachsignal entsteht, welches in einem Speicher gespeichert wird,
   - ein Faktor α definiert wird, um welchen das Sprachsignal verlängert oder verkürzt wird,
   - eine Fensterfunktion mit einem ersten steigenden Abschnitt der Länge N, einem zweiten, sich direkt an den ersten Abschnitt anschließenden, konstanten Abschnitt der Länge L und einem dritten, sich direkt an den zweiten Abschnitt anschließenden, fallenden Abschnitt definiert wird, wobei bei einer Überlagerung des ersten steigenden Abschnittes eines Fensters mit dem dritten fallenden Abschnitt eines anderen Fensters und einer Addition beider Abschnitte im Überlappungsbereich, sich das Ergebnis eins ergibt, was dem Wert des zweiten Abschnittes der Fensterfunktion entspricht,
   - aus dem digitalisierten, gespeicherten Sprachsignal in unregelmäßigen Abständen einer mittleren Länge αL

Segmente einer Länge L+N entnommen werden,

- diese, aus dem digitalisierten, gespeicherten Sprachsignal entnommenen, Segmente mit der Fensterfunktion im Zeitbereich gewichtet werden
- die gewichteten Segmente jeweils um eine definierte Anzahl von L Abtastwerten versetzt aufaddiert werden, wodurch das so entstehende Sprachsignal um den Faktor α verlängert bzw. um $1/\alpha$ verkürzt wird, **dadurch gekennzeichnet,**
- daß nacheinander an den Stellen der Entnahme der Segmente aus dem digitalisierten Sprachsignal, das dort entnommene, mit der Fensterfunktion gewichtete, Segment mit dem nachfolgend entnommenen, ebenfalls mit der Fensterfunktion gewichteten, Segment unter Ähnlichkeitsaspekten verglichen wird,
- daß zum schnellen Vergleich der Ähnlichkeit der Segmente lediglich der N Werte lange dritte, mit dem fallenden Fensterabschnitt gewichtete, Abschnitt des Segmentes mit dem jeweils ersten, mit dem steigenden N Werte langen Fensterabschnitt gewichteten Abschnitten des nachfolgenden Segmentes verglichen wird,
- daß diese Segmente zueinander versetzt aufaddiert werden, wenn die Ähnlichkeit beider verglichener Segmentteile maximal ist und
- daß zur Berechnung der Ähnlichkeit, als deren Maß, eine Korrelation verwendet wird.

2. Verfahren zur Geschwindigkeitsmodifikation von Sprachsignalen, insbesondere digitalisierten Sprachsignalen, bei dem

- ein analoges Sprachsignal digitalisiert wird, wodurch ein digitalisiertes Sprachsignal entsteht, welches in einem Speicher gespeichert wird,
- ein Faktor α definiert wird, um welchen das Sprachsignal verlängert oder verkürzt wird,
- eine Fensterfunktion mit einem ersten steigenden Abschnitt der Länge N, einem zweiten, sich direkt an den ersten Abschnitt anschließenden, konstanten Abschnitt der Länge L und einem dritten, sich direkt an den zweiten Abschnitt anschließenden, fallenden Abschnitt definiert wird, wobei bei einer Überlagerung des ersten steigenden Abschnittes eines Fensters mit dem dritten fallenden Abschnitt eines anderen Fensters und einer Addition beider Abschnitte im Überlappungsbereich, sich das Ergebnis eins ergibt, was dem Wert des zweiten Abschnittes der Fensterfunktion entspricht,
- aus dem digitalisierten, gespeicherten Sprachsignal in unregelmäßigen Abständen einer mittleren Länge αL Segmente einer Länge L+N entnommen werden,
- diese, aus dem digitalisierten, gespeicherten Sprachsignal entnommenen, Segmente mit der Fensterfunktion im Zeitbereich gewichtet werden,
- die gewichteten Segmente jeweils um eine definierte Anzahl von L Abtastwerten versetzt aufaddiert werden, wodurch das so entstehende Sprachsignal um den Faktor α verlängert bzw. um $1/\alpha$ verkürzt wird, **dadurch gekennzeichnet,**
- daß nacheinander an den Stellen der Entnahme der Segmente aus dem digitalisierten Sprachsignal, das dort entnommene Segment mit dem Resultat der Synthese mit dem nachfolgend entnommenen Segment verglichen wird,
- daß zum schnellen Vergleich der Abweichung des jeweiligen Syntheseresultats vom Originalsignal lediglich der N Werte lange dritte Abschnitt des zuletzt entnommenen Segmentes als Referenz herangezogen wird,
- daß diese Segmente zueinander versetzt aufaddiert werden, wenn die ermittelte Abweichung minimal ist und
- daß als Maß für die Abweichung der relative Fehler oder der absolute quadratische Fehler herangezogen wird.